Observations Thinning In Data Assimilation Computations

Gratton, Serge; Rincon-Camacho, Monserrat; Simon, Ehouarn; Toint, Ph

Published in:
EURO Journal on Computational Optimization

Publication date:
2015

Document Version
Early version, also known as pre-print

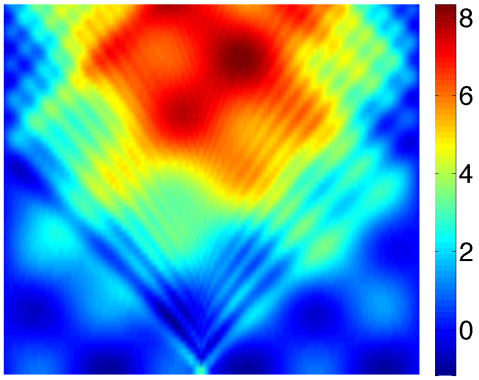
UNIVERSITÉ DE NAMUR

naxys
Namur Center for Complex Systems

Observation Thinning In Data Assimilation Computations

by S. Gratton, M. M. Rincon-Camacho, E. Simon and Ph. L. Toint

Report NAXYS-06-2013 30 November 2013



ENSEEIHT, 2, rue Camichel, 31000 Toulouse, France

CERFACS, 42, avenue Gaspard Coriolis, 31057 Toulouse, France

University of Namur, 61, rue de Bruxelles, B5000 Namur (Belgium)

http://www.unamur.be/sciences/naxys

Observation Thinning In Data Assimilation Computations

Serge Gratton[*], Monserrat Rincon-Camacho[†], Ehouarn Simon[‡], and Philippe L. Toint[§]

February 21, 2014

Abstract

We propose to use an observation-thinning method for the efficient numerical solution of large-scale incremental four dimensional (4D-Var) data assimilation problems. This decomposition is based on exploiting an adaptive hierarchy of the observations. Starting with a low-cardinality set and the solution of its corresponding optimization problem, observations are successively added based on a posteriori error estimates. The particular structure of the sequence of associated linear systems allows the use of a variant of the conjugate gradient algorithm which effectively exploits the fact that the number of observations is smaller than the size of the vector state in the 4D-Var model. The new algorithm is tested on a 1D-wave equation and on the Lorenz-96 system, the latter one being of special interest because of its similarity with Numerical Weather Prediction (NWP) systems.

Keywords: Data assimilation, numerical algorithms, multilevel optimization, a posteriori errors.

1 Introduction

Because of their ubiquitous application, data assimilation techniques for weather forecasting have been the subject of intensive study in the last decennies. In particular, dedicated numerical methods have been the subjet of much research : the large-dimensional nature and structure of the problem have prompted the proposal of a number of specialized algorithms, amongst which the influential and widely-used incremental 4D-VAR method (Courtier, Thépaut and Hollingsworth, 1994). In this method, a quadratically regularized optimization problem is solved, whose objective is to fit the modelled trajectory of the atmosphere'state to a potentially large set of observations. As most variants of the Gauss-Newton algorithm (see Gratton, Lawless and Nichols, 2007 for the connection with this classical tool), the 4D-VAR algorithm requires the sequential (possibly approximate) solution of linear least-squares subproblems. Each of these subproblems typically involves a number of variables sometimes considerably larger than the number of (linearized) observations. As a consequence, and in view of the very significant computational effort necessary for its solution, it is highly desirable to exploit the lower dimensionality of the observation space. Moreover, if it is at all possible to reduce this dimension by "thinning" the observation set, the benefit obtained by this dimension reduction is amplified.

In what follows, we present an algorithm for the solution of the 4D-VAR subproblem in which this cumulative advantage may be obtained. As will be described below, the goal of limiting the linear algebra computations to the lower-dimensional observation space can been achieved by using a "dual iterative method". The thinned observation set will be defined using a hierarchy of observations, from coarsest to finest level. Starting from the coarsest set of observations,

[*]ENSEEIHT, 2, rue Camichel, 31000 Toulouse, France and CERFACS, 42, avenue Gaspard Coriolis, 31057 Toulouse, France. (serge.gratton@enseeiht.fr).

[†]CERFACS, 42, avenue Gaspard Coriolis, 31057 Toulouse, France. (monserratrc@cerfacs.fr).

[‡]ENSEEIHT and IRIT, 2, rue Camichel, 31000 Toulouse, France, (ehouarn.simon@enseeiht.fr).

[§]Namur Center for Complex Systems, University of Namur, 61, rue de Bruxelles, B-5000 Namur, Belgium. (philippe.toint@unamur.be).

observations from the next level will be included in the observation set according to the influence they have on the solution, as measured by an estimate on the solution variation between two consecutive levels.

Our technique is similar in spirit to techniques well-known in the multigrid and mesh refinement literature (see Rincon-Camacho (2011), Brandt (1973) and McCormick (1984)). The main contribution of this paper is to rewrite this technique in a dual space associated with 4D-Var minimization problem (which is also the observation space), and to combine it with an efficient quadratic solver. Our approach notably differs from previous work (Daescu and Navon, 2004, or Cardinali, Pezzulli and Andersson, 2004, for instance) in the sense that it is not required to solve the problem on the finest level to select the influential observation but merely to compute residual estimate along the grid levels. The paper is structured as follows: Section 2 introduces the 4D-Var vocabulary and the considered hierarchy of observations. Section 3 then covers the associated error estimates and Section 4 presents the new algorithm. Section 5 discusses its application to a one-dimensional nonlinear wave equation and to the Lorenz96 model. Conclusions and perspectives are discussed in Section 6.

2 The problem and associated observation hierarchy

Consider the nonlinear least-squares problems in 4D-Var data assimilation, whose objective is to find an initial vector state at an initial time denoted as $x = x(t_0) \in \mathbb{R}^n$. The structure of the problem is as follows:

$$\min_{x\in\mathbb{R}^n} \tfrac{1}{2}\|x - x_b\|^2_{B^{-1}} + \tfrac{1}{2}\sum_{j=0}^{N_t} \|\mathcal{H}_j(x(t_j)) - y_j\|^2_{R_j^{-1}} \tag{2.1}$$

where the squared norm $\|x\|^2_M$ is induced by the inner product $x^T M x$ for a symmetric positive definite matrix $M \in \mathbb{R}^{l\times l}$ and a vector $x \in \mathbb{R}^l$. Here $x_b \in \mathbb{R}^n$ is the background vector, which is an a priori estimate. The vector $y_j \in \mathbb{R}^{m_j}$ is the vector of observations at time t_j and $\mathcal{H}_j$ is the operator modeling the observation process at the same time. The state vector $x(t_j)$ satisfies the nonlinear model of evolution $x(t_j) = \mathcal{M}_{0\to j}[x(t_0)]$. The matrix $B \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix representing the background-error covariance and the matrix $R_j \in \mathbb{R}^{m_j\times m_j}$ is also a symmetric positive definite matrix representing the observation-error covariance at time t_j.

The nonlinear least-squares problem (2.1) is solved iteratively. A possible approach could be based on a Newton method in which a quadratic approximation of the cost function would be computed using second order derivatives. However this approach is impractical for large scale problems, due to the cost of computing the quadratic models. Practical (first order) algorithms are rather based on a linearization of the nonlinear observation operator $\mathcal{H}_j(x(t_j))$ at the iterate x_k (which for the moment we denote only by x), leading to the optimization problem

$$\min_{\delta x\in\mathbb{R}^n} \tfrac{1}{2}\|x - x_b + \delta x\|^2_{B^{-1}} + \tfrac{1}{2}\|H\,\delta x - d\|^2_{R^{-1}} \tag{2.2}$$

where $R = \mathrm{diag}(R_0, R_1, \ldots, R_{N_t}) \in \mathbb{R}^{m\times m}$, $d \in \mathbb{R}^m$ denotes the concatenated misfits over time $y_j - \mathcal{H}_j(x(t_j))$ and $H \in \mathbb{R}^{m\times n}$ is the concatenated version of the linearized observation process

$$d = \begin{bmatrix} d_0 \\ d_1 \\ \vdots \\ d_{N_t} \end{bmatrix} \in \mathbb{R}^m, \quad H = \begin{bmatrix} H_0 \\ H_1 M_{0\to 1} \\ \vdots \\ H_{N_t} M_{0\to N_t} \end{bmatrix} \in \mathbb{R}^{n\times m}.$$

Here H_j and $M_{0\to j}$ are a (possibly approximate) linearization of the observation operator $\mathcal{H}_j$ and the model $\mathcal{M}_{0\to j}$ around $x = x_k(t_0)$ respectively.

Diverse techniques for solving problem (2.1) have been proposed, see for instance Courtier et al. (1994) for the so called incremental method which is equivalent to applying a truncated

Gauss-Newton iteration to problem (2.1) (see Gratton et al., 2007). The general algorithm is the following:

Algorithm 2.1 ***Incremental 4D-Var***

1. *Initialize* $x_0 = x_b \in \mathbb{R}^n$ *and set* $k = 0$.
2. *Compute* $x_k(t_j)$ *from* $x_k(t_0) = x_k$ *by running the nonlinear model* $\mathcal{M}$ *from time* t_0 *to* t_{N_t}.
3. *Calculate the vectors* $d_{k,j} = y_j - \mathcal{H}_j(x_k(t_j))$ *for* $j = 1, \ldots, N_t$.
4. *Find an approximate solution* δx_k *of the minimization problem (2.2), where* $x = x_k$, $H = H_k$ *and* $d = d_k$.
5. *Update the current solution as* $x_{k+1} = x_k + \delta x_k$.
6. *Set* $k := k + 1$. *If convergence is not achieved return to Step 2.*

This algorithm is known as the outer loop and the minimization step (Step 4) as the inner loop. Our interest is to reduce the cost of this optimization problem. Termination criteria for the outer and inner loops of this algorithm are discussed in Gratton et al. (2007).

It is most natural to solve the subproblem in Step 4 of Algorithm 2.1 directly in the space of dimension n, the size of x. This technique is referred to as the *primal approach*. At variance, the quadratic optimization problem (2.2) may be rewritten in a space of dimension m, the number of observations. This is known as the *dual approach* and it is especially useful when m is significantly smaller than n, in which case the second term in (2.2) is of relatively low rank compared to the first. A first approach of this type is the Physical-space Statistical Analysis System (PSAS) method (see Courtier, 1997), where the "low rank" observation term in (2.2) is handled by using a Sherman-Morrison formula and applying the standard conjugate-gradient algorithm (see Hestenes and Stiefel, 1952, or Golub and Van Loan, 1996, Section 10.2, p. 520) to the reduced system. This method has the drawback of not maintaining the inherent monotonicity of the conjugate-gradient algorithm in $\mathbb{R}^n$, thereby making any stopping rule of the inner minimization difficult to define (see El Akkroui, Gauthier, Pellerin and Buis, 2008, or Gratton, Gürol and Toint, 2010). Fortunately, a better alternative is available, where the conjugate-gradient algorithm is reformulated in $\mathbb{R}^m$ using the inner product defined by the metric HBH^T in order to guarantee the desired monotonicity properties without incurring additional cost. This technique, known as the Restricted Preconditioned Conjugate Gradient method (RPCG, see Gratton and Tshimanga, 2009 and Gratton et al., 2010), provides an efficient numerical procedure where substantial computing gains are obtained when $m \ll n$. We refer the reader to the cited publications for details.

Our aim in this work is to make the best possible use of this dual technique and to propose an adaptive observations' strategy for solving the data assimilation problem (2.1). Suppose that we have a large set of m observations $\mathcal{O}$ which can be decomposed into a hierarchical collection of sets $\{\mathcal{O}_i\}_{i=0}^r$, each with cardinality m_i, such that

$$\mathcal{O}_i \subset \mathcal{O}_{i+1} \quad \text{and} \quad m_i < m_{i+1} \quad (i = 0, \ldots, r-1)$$

where, by convention, $\mathcal{O}_r = \mathcal{O}$ and $m_r = m$. To each set of observations $\mathcal{O}_i$ we associate a misfit vector $y_i \in \mathbb{R}^{m_i}$ (by selecting the relevant components in the vector y), and the corresponding observation-error covariance matrix R_i. Given $\{\mathcal{O}_i\}_{i=0}^r$, we may therefore consider the hierarchical collection of minimization problems

$$\min_{x \in \mathbb{R}^n} \tfrac{1}{2}\|x - x_b\|^2_{B^{-1}} + \tfrac{1}{2}\|\mathcal{H}_i(x) - y_i\|^2_{R_i^{-1}}, \quad i = 0, \ldots, r, \tag{2.3}$$

where the vector $\mathcal{H}_i(x)$ denotes the nonlinear observation operator associated with the set of observations $\mathcal{O}_i$ concatenated over time.

Thus our objective is to construct the collection $\{\mathcal{O}_i\}_{i=0}^r$ such that the sequential solution of the problems (2.3) is obtained at significantly lower cost compared to solving (2.2) directly, while at the same time maintaining equivalent accuracy requirements.

3 Mathematical analysis

Our adaptive method is based on the exploitation of a posteriori bounds for the error between the solutions obtained using few observations or many. In particular, we are interested in comparing the accuracy of the solutions of problems (2.3) for $\mathcal{O}_i$ and $\mathcal{O}_{i+1}$. For simplicity of notation, we denote these sets as $\mathcal{O}_c$ a set with m_c observations and $\mathcal{O}_f$ a set containing m_f observations such that $m_c < m_f$ and $\mathcal{O}_c \subset \mathcal{O}_f$, where the indices c and f stand for 'coarse' and 'fine', respectively. The 'fine' optimization problem (2.3) is given, for the starting point x, by

$$\min_{\delta x_f \in \mathbb{R}^n} \tfrac{1}{2}\|x + \delta x_f - x_b\|^2_{B^{-1}} + \tfrac{1}{2}\|H_f \delta x - d_f\|^2_{R_f^{-1}} \tag{3.4}$$

where $d_f = y_f - \mathcal{H}_f(x)$ and H_f is the linearized version of the observation process at x involving the set of observations $\mathcal{O}_f$. We may reformulate (3.4) as a convex quadratic problem with linear equality constraints given by

$$\min_{\delta x_f \in \mathbb{R}^n,\, v_f \in \mathbb{R}^{m_f}} \tfrac{1}{2}\|x + \delta x_f - x_b\|^2_{B^{-1}} + \tfrac{1}{2}\|v_f\|^2_{R_f^{-1}}, \quad \text{subject to } \; v_f = H_f \delta x_f - d_f, \tag{3.5}$$

see Gratton et al. (2010). The Lagrange function corresponding to this reformulated problem is given by

$$\mathcal{L}(\delta x_f, v_f, \lambda_f) \stackrel{\rm def}{=} \tfrac{1}{2}\|x + \delta x_f - x_b\|^2_{B^{-1}} + \tfrac{1}{2}\|v_f\|^2_{R_f^{-1}} - \lambda_f^T (H_f \delta x_f - v_f - d_f).$$

Using this function, the optimality conditions for problem (3.5) can then be expressed as

$$\begin{aligned}
\nabla_{\delta x_f} \mathcal{L}(\delta x_f, v_f, \lambda_f) &= B^{-1}(x + \delta x_f - x_b) - H_f^T \lambda_f = 0, \\
\nabla_{v_f} \mathcal{L}(\delta x_f, v_f, \lambda_f) &= R_f^{-1}(v_f) + \lambda_f = 0, \\
\nabla_{\lambda_f} \mathcal{L}(\delta x_f, v_f, \lambda_f) &= v_f - H_f \delta x_f + d_f = 0,
\end{aligned}$$

which leads to the system

$$\begin{aligned}
H_f^T \lambda_f &= B^{-1}(x + \delta x_f - x_b), \\
-\lambda_f &= R_f^{-1}(H_f \delta x_f - d_f).
\end{aligned} \tag{3.6}$$

Gratton et al. (2010) and Gratton and Tshimanga (2009) show that the solution of this system can be obtained by solving

$$(R_f^{-1} H_f B H_f^T + I_{m_f})\lambda_f = R_f^{-1}(d_f - H_f(x_b - x)), \tag{3.7}$$

or, equivalently,

$$(H_f B H_f^T + R_f)\lambda_f = d_f - H_f(x_b - x), \tag{3.8}$$

for λ_f and then substituting in the second equation of (3.6) for δx_f and v_f.

We now compare the solution $(\delta x_f, \lambda_f)$ to the solution of the coarse level minimization problem

$$\min_{\delta x_c \in \mathbb{R}^n} \tfrac{1}{2}\|x + \delta x_c - x_b\|^2_{B^{-1}} + \tfrac{1}{2}\|\Gamma_f(H_f \delta x_c - d_f)\|^2_{R_c^{-1}}, \tag{3.9}$$

where Γ_f is the restriction operator $\Gamma_f : \mathbb{R}^{m_f} \to \mathbb{R}^{m_c}$ from the fine observation space to the coarse one. For the purpose of the calculus, let Π_c be the prolongation operator from the coarse observation space to the fine one such as

$$\Pi_c \stackrel{\rm def}{=} \sigma_f \Gamma_f^T \tag{3.10}$$

for some $\sigma_f > 0$. The coarse level minimization problem reads

$$\min_{\delta x_c \in \mathbb{R}^n} \tfrac{1}{2}\|x + \delta x_c - x_b\|^2_{B^{-1}} + \tfrac{1}{2}\|\Pi_c^T(H_f \delta x_c - d_f)\|^2_{R_c^{-1}}, \tag{3.11}$$

with $\bar{R}_c^{-1} = (\frac{1}{\sigma_f})^2 R_c^{-1}$. As above, we reformulate (3.11) as a convex quadratic problem with linear equality constraints, and obtain

$$\min_{\delta x_c \in \mathbb{R}^n,\, v_c \in \mathbb{R}^{m_c}} \tfrac{1}{2}\|x + \delta x_c - x_b\|^2_{B^{-1}} + \tfrac{1}{2}\|v_c\|^2_{\bar{R}_c^{-1}}, \quad \text{subject to } v_c = \Pi_c^T(H_f \delta x_c - d_f), \tag{3.12}$$

whose Lagrangian function is given by

$$\mathcal{L}(\delta x_c, v_c, \lambda_c) \stackrel{\rm def}{=} \tfrac{1}{2}\|x + \delta x_c - x_b\|^2_{B^{-1}} + \tfrac{1}{2}\|v_c\|^2_{\bar{R}_c^{-1}} - \lambda_c^T(\Pi_c^T H_f \delta x_c - v_c - \Pi_c^T d_f).$$

The optimality conditions now become

$$\begin{array}{rcl} H_f^T \Pi_c \lambda_c & = & B^{-1}(x + \delta x_c - x_b), \\ -\lambda_c & = & \bar{R}_c^{-1}\Pi_c^T(H_f \delta x_c - d_f) \end{array} \tag{3.13}$$

These conditions may again be solved by computing λ_c such that

$$(\bar{R}_c^{-1}\Pi_c^T H_f B H_f^T \Pi_c + I_{m_c})\lambda_c = \bar{R}_c^{-1}\Pi_c^T(d_f - H_f(x_b - x)). \tag{3.14}$$

Since the dimension of the system (3.14) is smaller than that of the system (3.7), solving problem (3.11) is (often much) cheaper than solving problem (3.4). The RPCG algorithm mentioned in the previous section derives its efficiency by using the formulation (3.14) rather than (3.7), see Gratton et al. (2010) and Gratton and Tshimanga (2009). Note that both algorithms (conjugate gradient on (3.4) preconditioned by B or RPCG) generate the same iterates in exact arithmetic and that the main gain is obtain by the fact that RPCG explores the structure of a smaller quadratic optimization problem.

After obtaining the solution $(\delta x_c, \lambda_c)$, our objective is now to compute the difference between the (still unknown) λ_f and $\Pi_c\lambda_c$ in order to identify the set of observations where this difference is large. Our intention is then to construct the "fine" problem from the "coarse" one by adding to the coarse the observations that are singled out by this comparison. Our motivations for building a criterion based on the Lagrange multipliers λ are twofold. First, they provide information on the variations of the cost function due to perturbations in the right hand side of the constraint, e.g. changes in the observation values or network: for a small perturbation ϵ, one has that $J(x(\epsilon)) = J(x(0)) - \lambda^T\epsilon + \mathcal{O}(\|\epsilon\|^2)$ - see Nocedal and Wright (1999) for instance. Secondly, they are directly linked to the primal solution thanks to the operator BH^T: $\delta x = x^b - x + BH^T\lambda$. Thus, the norm (associated to some positive definite matrix) of the difference between the theoretical optimal increment δx of the primal problem and an approximation $\delta\tilde{x}$ is nothing more than the norm, associated to a different positive definite matrix, of the difference between the theoretical optimal Lagrange multipliers λ and its approximation $\tilde{\lambda}$: $\|\delta x - \delta\tilde{x}\|_M = \|\lambda - \tilde{\lambda}\|_{HBMBH^T}$.

We start by computing the difference between λ_f and $\Pi_c\lambda_c$ in the energy norm $\|.\|_{H_f B H_f^T + R_f}$ associated with the system (3.8), separating the desired expression in two terms. More precisely, if λ is the exact solution of the quadratic problem, the k-th iterate of conjugate gradient minimizes the distance to λ in this energy norm. This norm is therefore often used when analyzing the conjugate gradient method (see Arioli (2004) for instance). Using (3.6) and (3.13), we define

$$\begin{array}{rcl} E_1 & \stackrel{\rm def}{=} & \|\lambda_f - \Pi_c\lambda_c\|^2_{R_f} \\ & = & \langle R_f\left(\lambda_f - \Pi_c\lambda_c\right), -R_f^{-1}(H_f\delta x_f - d_f) - \Pi_c\lambda_c\rangle \\ & = & \langle \lambda_f - \Pi_c\lambda_c, -H_f\delta x_f + d_f - R_f\Pi_c\lambda_c\rangle, \end{array}$$

and

$$\begin{array}{rcl} E_2 & \stackrel{\rm def}{=} & \|\lambda_f - \Pi_c\lambda_c\|^2_{H_f B H_f^T} \\ & = & \|H_f^T\lambda_f - H_f^T\Pi_c\lambda_c\|^2_B \\ & = & \langle B(H_f^T\lambda_f - H_f^T\Pi_c\lambda_c), B^{-1}(x + \delta x_f - x_b) - H_c^T\Pi_c\lambda_c\rangle \\ & = & \langle \lambda_f - \Pi_c\lambda_c, H_f(x - x_b) + H_f\delta x_f - H_f B H_f^T\Pi_c\lambda_c\rangle. \end{array}$$

By adding these errors E_1 and E_2 we obtain that

$$\begin{aligned} E_1 + E_2 &= \langle \lambda_f - \Pi_c\lambda_c, -H_f\delta x_f + d_f - R_f\Pi_c\lambda_c\rangle \\ &\quad + \langle \lambda_f - \Pi_c\lambda_c, H_f(x - x_b) + H_f\delta x_f - H_fBH_f^T\Pi_c\lambda_c\rangle \\ &\quad + \langle \lambda_f - \Pi_c\lambda_c, H_f\delta x_c - H_f\delta x_c\rangle \\ &= \langle \lambda_f - \Pi_c\lambda_c, d_f - R_f\Pi_c\lambda_c - H_f\delta x_c\rangle \\ &\quad + \langle \lambda_f - \Pi_c\lambda_c, H_f(x - x_b + \delta x_c) - H_fBH_f^T\Pi_c\lambda_c\rangle. \end{aligned}$$

But the first equation of (3.13) multiplied by H_fB gives that

$$H_fBH_f^T\Pi_c\lambda_c = H_f(x - x_b + \delta x_c), \tag{3.15}$$

and hence that

$$\langle \lambda_f - \Pi_c\lambda_c, H_f(x - x_b + \delta x_c) - H_fBH_f^T\Pi_c\lambda_c\rangle = 0 \tag{3.16}$$

Therefore, one has

$$E_1 + E_2 = \langle \lambda_f - \Pi_c\lambda_c, d_f - H_f\delta x_c - R_f\Pi_c\lambda_c\rangle \tag{3.17}$$

Let M be a positive definite matrix. The equation (3.17) can read

$$E_1 + E_2 = \langle M^{1/2}(\lambda_f - \Pi_c\lambda_c), M^{-1/2}(d_f - H_f\delta x_c - R_f\Pi_c\lambda_c)\rangle$$

Thus, we obtain

$$E_1 + E_2 \leq \|\lambda_f - \Pi_c\lambda_c\|_M \|d_f - H_f\delta x_c - R_f\Pi_c\lambda_c)\|_{M^{-1}}$$

Furthermore, one has

$$E_1 + E_2 = \|\lambda_f - \Pi_c\lambda_c\|^2_{R_f + H_fB_fH_f^T} \tag{3.18}$$

Thus, choosing M equal to the matrix $R_f + H_fB_fH_f^T$, the previous inequality leads to the a posteriori error bound

$$(E_1 + E_2)^{1/2} \leq \|d_f - H_f\delta x_c - R_f\Pi_c\lambda_c\|_{(R_f + H_fBH_f^T)^{-1}}$$

or equivalently

$$E_1 + E_2 \leq \|d_f - H_f\delta x_c - R_f\Pi_c\lambda_c\|^2_{(R_f + H_fBH_f^T)^{-1}}$$

However, the computation of the inverse of $R_f + H_fBH_f^T$ is not an easy task in the variational data assimilation framework due to the nature of B (complex matrix-vector operator) and can be very expensive. If one rather considers using $M = R_f$, we obtain that

$$E_1 + E_2 \leq \|\lambda_f - \Pi_c\lambda_c\|_{R_f} \|d_f - H_f\delta x_c - R_f\Pi_c\lambda_c)\|_{R_f^{-1}}$$

Because R_f and $H_fBH_f^T$ are positive semi-definite matrices, one has

$$\|\lambda_f - \Pi_c\lambda_c\|^2_{R_f} \leq \|\lambda_f - \Pi_c\lambda_c\|^2_{R_f + H_fBH_f^T}$$

resulting in

$$E_1 + E_2 \leq \|\lambda_f - \Pi_c\lambda_c\|_{R_f + H_fBH_f^T} \|d_f - H_f\delta x_c - R_f\Pi_c\lambda_c)\|_{R_f^{-1}}$$

and thus, using (3.18),

$$E_1 + E_2 \leq \|d_f - H_f\delta x_c - R_f\Pi_c\lambda_c\|^2_{R_f^{-1}} \tag{3.19}$$

As a consequence, the left-hand side of this inequality (the sought error estimate) can be bounded above using inequality (3.19), giving the following a posteriori error.

Theorem 3.1 *Let δx_f be the solution to the problem (3.5) and λ_f the corresponding Lagrange multiplier to the constraint in (3.5) such that the couple (δx_f, λ_f) satisfies (3.6). Analogously, let δx_c be the solution to (3.12) and λ_c the corresponding Lagrange multiplier such that (δx_c, λ_c) satisfies (3.13). Then the a posteriori error bound satisfies the inequality*

$$\|\lambda_f - \Pi_c \lambda_c\|^2_{R_f + H_f B H_f^T} \quad \leq \quad \|d_f - H_f \delta x_c - R_f \Pi_c \lambda_c\|^2_{R_f^{-1}} \tag{3.20}$$

Note that the bound (3.20) does not involve the computation of λ_f or δx_f.

The use of an a posteriori error bound depending on a primal-dual problem in adaptive finite elements was studied in Section 12.2 (p. 100) of Rincon-Camacho (2011).

In the next section we describe how to make use of the bound (3.20) in order to algorithmically construct the 'fine' problem from the 'coarse' one.

4 A new adaptive algorithm

Starting from a small set of observations $\mathcal{O}_c$, our goal is to add only significant observations to produce $\mathcal{O}_f$ so that the a posteriori error (3.20) is reduced. Our strategy is to define an auxiliary set of *potential* fine observations $\tilde{\mathcal{O}}_f$ from which the observations in $\mathcal{O}_f$ are selected. However, describing our strategy (and algorithm) requires additional assumptions on the hierarchy of (potential) observations. More specifically, we complete our assumptions as follows.

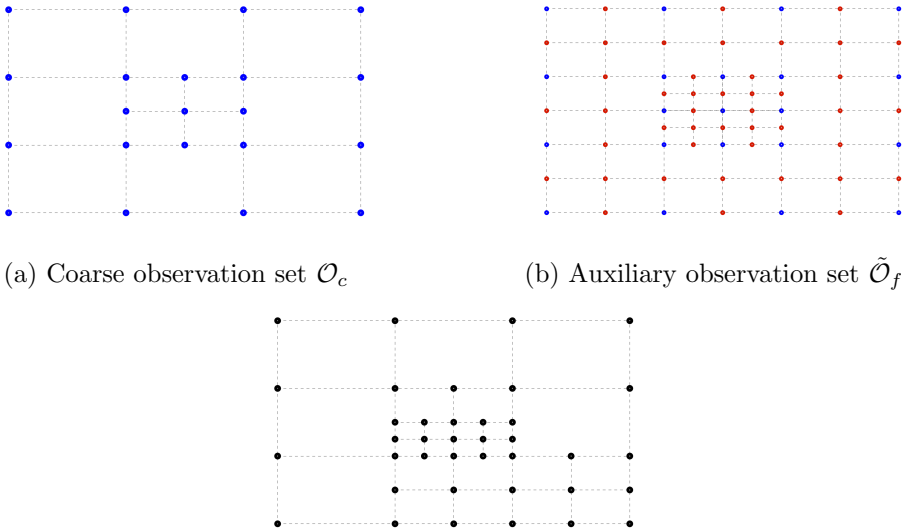
- The observations correspond to localizations in some underlying continuous measurable "observation space".
- The coarse observation set partitions the observation space in a finite number of *cells* $\{c_j\}_{j=1}^{p_c}$ of measures $\{w_j\}_{j=1}^{p_c}$.
- The auxiliary set $\tilde{\mathcal{O}}_f$ is constructed by considering all observations in $\mathcal{O}_c$ with the addition of a single additional potential observation point in the interior of each cell. The cell is said to be associated with this additional potential observation.
- There exists a prolongation operator $\tilde{\Pi}_c$ from $\mathcal{O}_c$ to $\tilde{\mathcal{O}}_f$ such that, for each potential observation o_j in $\tilde{\mathcal{O}}_f \setminus \mathcal{O}_c$, $\tilde{\Pi}_c$ defines the value of this observation only in terms of the observations of the associated cell c_j. As expected, we define $\tilde{\Pi}_c = \sigma_f \tilde{\Gamma}_f^T$.

We illustrate these assumptions by a 2D example: suppose that the observation space is the plane and the coarse observation set $\mathcal{O}_c$ is the rectangular 'grid' shown in Figure 4.1 (a): the cells are elementary rectangles in this grid, whose measure is given by their surface. We may then define $\tilde{\mathcal{O}}_f$ as the grid shown in Figure 4.1 (b), which we obtained by locally adding a new potential observation in the center of each rectangle and four additional ones on its boundary (effectively doubling the mesh in every direction). The observations in $\mathcal{O}_f$ (as shown in Figure 4.1 (c)) can then be extracted from $\tilde{\mathcal{O}}_f$.

In this example, the restriction operator $\tilde{\Gamma}_f$ can be defined as the usual full weighting operator associated with bilinear interpolation prolongations (which justifies the introduction of the boundary points). This full-weighting restriction operator is given, on every grid node, by the stencil

$$\frac{1}{16} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}. \tag{4.21}$$

We presented our approach on a 2D example, that is well suited to observation on the surface of the Earth. Note that this technique would also apply for other observation types, such as satellite along-track data. In this case, the observations can be located along track of a satellite, in which case the cells could be arcs located on the track. Note also that for convenience of



(a) Coarse observation set $\mathcal{O}_c$

(b) Auxiliary observation set $\tilde{\mathcal{O}}_f$

(c) Fine observation set $\mathcal{O}_f$

Figure 4.1: Auxiliary observations

the representation, we considered regularly spaced observations in picture 4.1. This is not a requirement of the method. All what we need is a hierarchy involving cells in the observation set.

In order to achieve our goal to select 'important' observations from $\tilde{\mathcal{O}}_f$, we need to compute localized error indicators (3.20). We define them as

$$\eta_j \stackrel{\rm def}{=} w_j \, \langle (\tilde{d}_f - \tilde{H}_f \delta x_c - \tilde{R}_f \tilde{\Pi}_c \lambda_c)|_j, (\tilde{R}_f^{-1}(\tilde{d}_f - \tilde{H}_f \delta x_c - \tilde{R}_f \tilde{\Pi}_c \lambda_c))|_j \rangle,$$

where $\tilde{d}_f$, $\tilde{H}_f$, $\tilde{R}_f$ are constructed from the set of observations $\tilde{O}_f$ and where the symbol $|_j$ denotes the restriction of the associated quantity to the cell c_j. Note that the η's correspond to the term on the right-hand side of (3.20). This bound was obtained by solving the optimization problem (3.11) in which the original cost function has be rewritten in order to involve the transpose of Π_c the prolongation operator and the modified observation error covariance matrix rather than the restriction operator Γ_f. It does not impact the optimal increment δx_c because the optimization problems (3.9) and (3.11) are equivalent. However, the Lagrange multiplier λ_c obtained solving (3.11) is equal to the one obtained solving (3.9) scaled by $\frac{1}{\sigma_f}$.

In order to decide which cells will be chosen to include a new interior potential observation, we use the bulk chasing strategy (also known as Dörfler marking, see, for instance, Dörfler, 1996, Morin, Nochetto and Siebert, 2000, Logg, Mardal and Wells, 2012). For a constant $\theta_1 \in (0,1)$, we construct a minimal set $\mathcal{S}_\eta$ such that

$$\theta_1 \left(\sum_{j=1}^{p} \eta_j \right) \leq \sum_{k \in \mathcal{S}_\eta} \eta_k \tag{4.22}$$

where p is the amount of observations in $\tilde{\mathcal{O}}_f$. In practice this construction is carried out by progressively constructing each set using a greedy heuristic which includes first the non-included cell with maximal indicator value.

Once $\mathcal{S}_\eta$ is constructed, we decide that a cell k of $\mathcal{O}_c$ is 'refined' if $k \in \mathcal{S}_\eta$, meaning that the observations associated with the corresponding cell in $\tilde{\mathcal{O}}_f$ are added to the set $\mathcal{O}_c$ to construct the

new set O_f. More formally,

$$\mathcal{O}_f \stackrel{\rm def}{=} \mathcal{O}_c \cup \left(\bigcup_{k \in \mathcal{S}_\eta} o_k \right). \tag{4.23}$$

Thus, starting with a small set of observations $\mathcal{O}_0$, we progressively add observations using the method just described, resulting in the following algorithm.

Algorithm 4.1 ***An algorithm using adaptive observations***

1. *Set $i = 0$, initialize x and the coarse observation set $\mathcal{O}_0$.*
2. *Find the solution $(\delta x_i, \lambda_i)$ to the problem*

$$\min_{\delta x_i \in I\!R^n} \tfrac{1}{2}\|x_i + \delta x_i - x_b\|^2_{B^{-1}} + \tfrac{1}{2}\|H_i \delta x_i - d_i\|^2_{R_i^{-1}}, \tag{4.24}$$

   by approximately solving the system

$$(R_i^{-1} H_i B H_i^T + I_{m_i})\lambda_i = R_i^{-1}(d_i - H_i(x_b - x_i)) \tag{4.25}$$

   using RPCG and then setting $\delta x_i = x_b - x_i + B H_i^T \lambda_i$.
3. *Given the set of observations $\mathcal{O}_i$, construct the auxiliary set $\tilde{\mathcal{O}}_{i+1}$ such that the conditions described at the beginning of this section hold.*
4. *For each cell c_j of observation set $\mathcal{O}_i$ compute the error indicators*

$$\eta_j \;=\; w_j \, \langle (\tilde{d}_{i+1} - \tilde{H}_{i+1}\delta x_i - \tilde{R}_{i+1}\tilde{\Pi}_i\tilde{\lambda}_i)|_j, (\tilde{R}_{i+1}^{-1}(\tilde{d}_{i+1} - \tilde{H}_{i+1}\delta x_i - \tilde{R}_{i+1}\tilde{\Pi}_i\tilde{\lambda}_i))|_j \rangle$$

   with $\tilde{\lambda}_i$ a modified Lagrange multiplier.
5. *Build the set $\mathcal{S}_\eta$ such that*

$$\theta_1 \left(\sum_{j=1}^{p_{i+1}} \eta_j \right) \leq \sum_{k \in \mathcal{S}_\eta} \eta_k$$

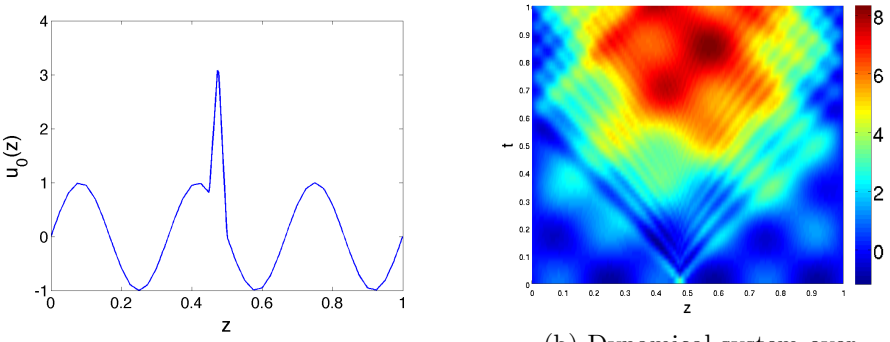
   using the bulk chasing strategy.
6. *Construct the set O_{i+1} as*

$$O_{i+1} := O_i \cup \left(\bigcup_{k \in \mathcal{S}_\eta} o_k \right)$$

7. *Update $x \leftarrow x + \delta x_i$.*
8. *Increment i and return to Step 2.*

We note that the computation of $(\delta x_i, \lambda_i)$ in Step 2 corresponds to applying the RCPG algorithm to (4.24), thereby making this computation essentially dependent on m, the number of observations, which the algorithm maintains as small as necessary by design.

Furthermore, we introduced a modified Lagrange multiplier $\tilde{\lambda}_i$ in Step 4 because λ_i is obtained solving (4.24) which corresponds to the original coarse resolution problem (3.9) and not (3.11). At convergence of the RPCG algorithm, one has that $\tilde{\lambda}_i = \frac{1}{\sigma_f}\lambda_i$, and so, a simple scaling of the Lagrange multiplier λ_i has to be done before computing the error estimate. However, we might prefer to approximately solve (4.25) in order to reduce the computational costs of the assimilation. In that case, the scaling factor $\frac{1}{\sigma_f}$ has to be properly introduced in the definition of the cost function and the algorithm RPCG. Both strategies - scaling the final Lagrange multiplier



(a) Initial $x_0 := u_0(z)$

(b) Dynamical system over space and time

Figure 5.2: Nonlinear 1D Wave equation

or introducing the scaling factor in the cost function and RPCG - led to similar results in our numerical experiments.

We also note that the bulk of the remaining work is in the computation of the error indicators in Step 4, as this requires the product of $\tilde{H}_{i+1}$ with δx_i.

Similar methods for constructing adaptive grids are the multi-level adaptive technique (MLAT) studied in Brandt (1973) and the fast adaptive composite grid (FAC) presented in McCormick (1984).

5 Numerical experiments

5.1 Two test problems

This section is devoted to showing the performance of our new adaptive algorithm on two test cases. The first one is a one-dimensional wave equation system, which we refer to as the 1D-Wave model from now on. The dynamics on this model are governed by the following nonlinear wave equations:

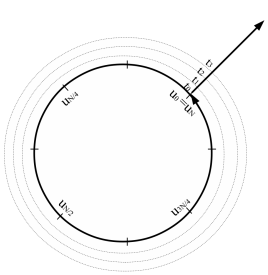
$$\begin{aligned} &\frac{\partial^2}{\partial t^2}u(z,t) - \frac{\partial^2}{\partial z^2}u(z,t) + f(u) = 0, \\ &u(0,t) = u(1,t) = 0, \\ &u(z,0) = u_0(z),\ \frac{\partial}{\partial t}u(z,0) = 0, \\ &0 \le t \le T,\ \ 0 \le z \le 1, \end{aligned} \tag{5.26}$$

where we have chosen $f(u) = \mu e^{\eta u}$. The spatial discretization involves 360 grid points, resulting in $\Delta x \approx 2.8 \cdot 10^{-3}$. We also set $T = 1$ and $\Delta t = \frac{1}{64}$. In this case we look for the initial function $u_0(z)$, which corresponds to x in the data assimilation problem (2.1). We illustrate in Figure 5.2 (a) the initial state vector u_0 ($x = u_0$) and the evolution of the system in Figure 5.2 (b) (view from the top) where the space domain corresponds to the horizontal axis and the time domain to the vertical axis.

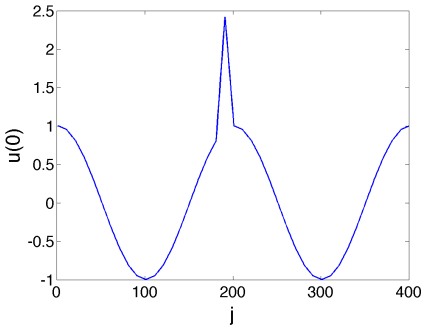
Our second example is the model referred as Lorenz96 presented in Lorenz and Emanuel (1998). The variable $\bar{u}$ is a vector of N-equally spaced entries around a circle of constant latitude, i.e. $\bar{u}(t) = (u_1(t), u_2(t), \ldots, u_N(t))$. The N-dimensional system is determined by the following N equations

$$\frac{du_j}{dt} = \frac{1}{\kappa}(-u_{j-2}u_{j-1} + u_{j-1}u_{j+1} - u_j + F), \quad j = 1, \ldots N, \tag{5.27}$$

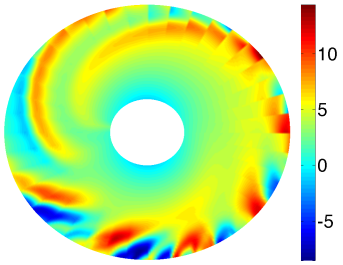
where F and κ are constants independent of j. To form a cyclic chain, we set $u_N = u_0$, $u_{-1} = u_{N-1}$ and $u_{N+1} = u_1$. This system is known to have a chaotic behaviour over time depending on the



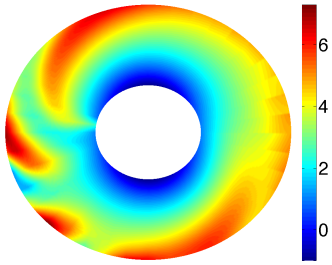
(a) Coordinate system



(b) Initial $x_0 := (u_1(0), u_2(0), \ldots, u_N(0))$



(c) Dynamical system over space and time



(d) Window of assimilation

Figure 5.3: Lorenz96 problem

parameters N, F and κ (see for instance Karimi and Paul, 2009) For a given set of parameters N, F and κ for which a stable behavior is observed, i.e. where data assimilation can be performed, we consider then the following dynamical system:

$$\frac{du_{j+\theta}}{dt} = \frac{1}{\kappa}(-u_{j+\theta-2}u_{j+\theta-1} + u_{j+\theta-1}u_{j+\theta+1} - u_{j+\theta} + F), \quad j = 1, \ldots N, \quad \theta = 1, \ldots, \Theta, \tag{5.28}$$

where θ and Θ are integers. Thus, the new size of the vector $\bar{u}(t)$ is $N \times \Theta$, which may be specified as large as needed in our numerical experiments. The dynamical system is plotted for $N = 40$, $F = 8$, $\kappa = 120$ and $\Theta = 10$, using the coordinate graph described in Figure 5.3 (a). For an initial state $x = \bar{u}(0)$ as that shown in Figure 5.3 (b) the system develops over time as described in Figure 5.3 (c) (view from the top). As we observe that the system becomes chaotic after a certain time, we consider a reduced window of assimilation plotted in Figure 5.3 (d). The time length of the assimilation window is $T = 120$ and $\Delta t = \frac{1}{80}$.

Twin experiments are performed in order to assess the performances of the suggested algorithm: a true state or reference simulation is built by running the model from a given initial condition. In both cases, the background is built from the initial true state by adding a noise following the normal distribution $\mathcal{N}(0, \sigma_b^2)$. In the same way, the observations are generated by adding to the true state a noise following the normal distribution $\mathcal{N}(0, \sigma_o^2)$. We chosen $\sigma_b = 0.2$ and $\sigma_o = 0.05$ in the wave1D model, and $\sigma_b = 0.2$ and $\sigma_o = 0.1$ in the Lorenz96 system. The background and observation error covariances matrices are assumed to be diagonal: $B = \sigma_b^2 I_n$ and $R = \sigma_o^2 I_p$, with σ_b and σ_o the variances of the normally distributed noise present in the background and the observations respectively, and I_n and I_p the identity matrices of dimension n and p.

5.2 Results

We now provide numerical results using Algorithm 4.1, which uses the RPCG algorithm in Step 3, as was mentioned in Section 2. When tuning the accuracy parameter for stopping the inner iterations, we noticed that it was suitable to choose a value in the middle of the log-scale range. In the case of the 1DWave equation the range of the parameter is more or less defined by $[10^{-4}, 10^{-1}]$, thus we choose 10^{-2}, while the range is approximately $[10^{-6}, 10^{-1}]$ in the case of the Lorenz96 model, and we chose the value 10^{-4} in our experiment.

For the 1DWave problem, we depict the background vector in a black dashed line together with the true solution as a red line, in Figure 6.4 (a). The result given by our algorithm is plotted as a blue line in Figure 6.4 (b). Both lines are almost indistinguishable highlighting the good performances of the algorithm to retrieve the true solution. The background vector and the true solution are plotted in Figure 6.5 (a) and the result given by our algorithm is presented in Figure 6.5 (b) for the Lorenz96 model.

In order to observe the adaptive nature of the algorithm, for an intermediate iteration i, we display two consecutive sets of observations $\mathcal{O}_i$ and $\mathcal{O}_{i+1}$ in Figures 6.6 (a)-(b) and 6.7 (a)-(b) for the 1DWave and the Lorenz96 respectively. In order to appreciate the impact of the local error indicators defined in Step 4, we also illustrate, in Figures 6.6 (d) and 6.7 (e), the local behaviour of the error between the prolongation of the current λ_i to the set $\mathcal{O}_{i+1}$ and the true $\tilde{\lambda}_{i+1}$, as given by

$$\epsilon_j = w_j \left\langle (\tilde{\lambda}_{i+1} - \Pi_i \lambda_i)|_j, [(\tilde{R}_{i+1} + \tilde{H}_{i+1} B \tilde{H}_{i+1}^T)(\tilde{\lambda}_{i+1} - \Pi_i \lambda_i)]|_j \right\rangle,$$

together with that of the local error indicator itself (η_j is displayed Figures 6.6 (c) and 6.7 (c)).

In the case of the 1DWave equation we observe in Figure 5.2 (a) that the peak in the middle of the signal produces a dynamical reaction over space and time, shown in part (b) of the same figure. The error η_j in Figure 6.6 (c) is larger in the regions of strong dynamical activity, whose identification is clear in Figure 6.6 (a)-(b) which shows the evolution of the observation sets. At this level i, large values of η_j are also present at the bottom of (c) resulting in the selection of observations at the right boundary of the spatial domain. In Figure 6.6 (d), we also observe that the regions where the difference ϵ_j is large coincide with those where η_j is also large in Figure 6.6 (c). However, the error η_j strongly overestimates the difference ϵ_j as noted by the large difference in amplitude between these two quantities.

In the case of the Lorenz96 model where the initial signal has also a peak in the middle (Figure 5.3 (b)), the high dynamics are on the edge of the space and time graph (Figure 5.3 (d)). In this case the quantity η_j also provides indication of where more observations are needed. In Figure 6.7 (c), it corresponds to the spatial area located at the edge of the outer circle (end of the assimilation window) and on the bottom of the domain during the second half of the simulation, for which the high dynamics is poorly represented. We also note that the distance ϵ_j in Figure 6.7 (d) is quite consistent with η_j even if the larger values of ϵ_j tend to be more located in the left side of the domain.

In Figures 6.8 and 6.9, we compare the performance of the new algorithm with the simple use of uniform observations and a benchmark method where a pre-established hierarchy of uniform distributed and progressively denser observations sets is used in Algorithm 4.1 (skipping Steps 4-6). This last method bypasses the new adaptive features of the method and its associated computing cost. For this comparison, we define, for a given iterate x resulting from Step 7 of Algorithm 4.1, the cost function as the value of the function in (2.3) when $i = r$, i.e. when all the possible observations are used. We then plot the evolution of this cost function (in logarithmic scale) against the number of observations used and the associated flop (floating-point operations) counts for the three algorithms. The curves for the uniform case do not correspond to an algorithm, but show the accuracy and computational costs associated with directly solving the problem on a uniform grid for each specific size.

We observe that the new method achieves a smaller cost function (plotted in red) than that obtained by using uniformly distributed observations (plotted in blue) in both cases, or the benchmark (plotted in black) in the Lorenz96 system. In this case, the selection of the observations

leads to a faster decrease of the cost function against the number of observations and computing costs. So, the final cost function achieved by the benchmark is obtained by the new algorithm assimilating 10 time less observations and for a computing cost divided by 10. In the 1DWave equation, similar cost functions are achieved by the benchmark and the new algorithm. However, the new algorithm is almost 50% cheaper than the benchmark in terms of the number of assimilated observations and computing costs. This is explained by the fact that most observations used by the adaptive algorithm are in regions of space where their contribution to accuracy is largest.

Finally, we plot the evolution with time of the Root Mean Square (RMS) error of the optimized solutions obtained during the last minimization:

$$RMS(t) = \frac{1}{n}\sqrt{\sum_{k=1}^{n}(x^t(t,k) - x(t,k))^2}$$

where t is the time, n is the dimension of the state vector, and x^t and x are the true and optimized solutions. The magenta line corresponds to the RMS error of the background, which means the solution without assimilating data. For the 1DWave equation, we note that both the benchmark and adaptive algorithms lead to the best solutions. For the Lorenz96 system, the best solution is obtained with the new algorithm for which the smallest cost function is achieved due to a fastest decrease. It leads to a better control of the error growth of the solution over the assimilation window.

Computational experience not reported here also indicates that the effect of the choice of starting value of x does not affect significantly the hierarchy of observation sets beyond the fact that observations concentrate in regions of high dynamical activity. We also found that $[0.4, 0.7]$ appears to be an adequate range for the parameter θ_1 in (4.22), yielding a satisfactory rate of inclusion of new observations at each iteration.

Both results on accuracy and computational costs are therefore highly encouraging.

6 Conclusions and perspectives

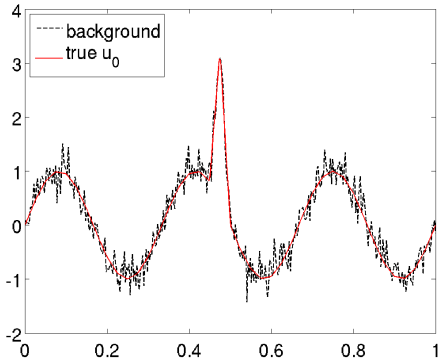
An algorithm for the solution of the 4D-VAR problem is proposed, which identifies the influential data and exploits this identification to improve on computing efficiency. The cpu-time gains are obtained for two cumulative reasons, the first being that the available number of observations is used very effectively, and the second the fact that the cost of the subproblem solution is significantly reduced by the use of dual-space conjugate-gradient techniques like RPCG. Numerical experience has been presented on two nonlinear test problems, and the results are encouraging.

Further refinements of the algorithm could be considered, such as the use of adaptive preconditioners in the subproblem solver, and continued experience with the method is of course desirable to assess its true potential. Extensions of these ideas in other domains are also possible: we think in particular of data assimilation problems in frameworks where each state of the system is itself an image on which adaptive reconstruction techniques could be applied. Moreover, our analysis relies on upper-bounds on the error on Lagrange multipliers associated with the observation in the 4D-Var minimization problem. These upper-bounds are based on a repeated use of the triangular and Cauchy-Schwarz inequalities. Our paper shows that based on these techniques, we can obtain a very efficient scheme to dynamically add observations in the course of the minimization. Further work in other Data Assimilation platforms would be needed to explore the relevance of the approach in other settings.

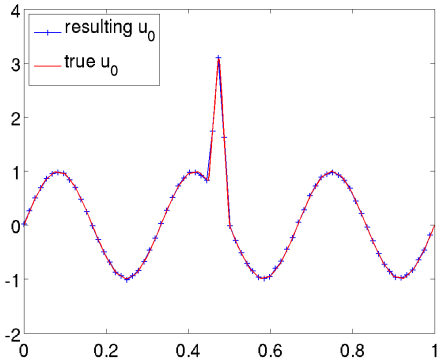
The authors believe that the hierarchy of observations described here may also be of interest in the more general framework of designing adaptive observation strategies. Indeed the new method proposed is able to isolate important observations from less important ones at a computational cost which is less than as single solution of the 4D-VAR problem involving all observations at the fine level, a unique feature to the authors' knowledge. This line of development is the object of ongoing research.

Acknowledgments.

The authors would like to thank S. Gürol and X. Vasseur at CERFACS for helpful scientific discussions.

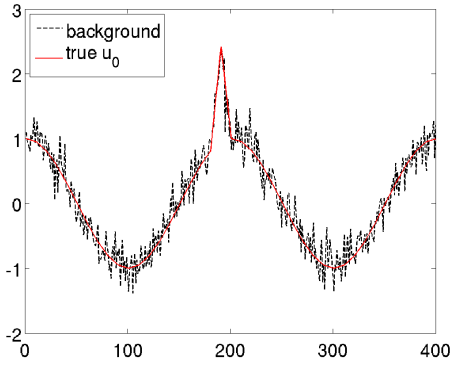


(a) Background vector
and true initial $u(0)$

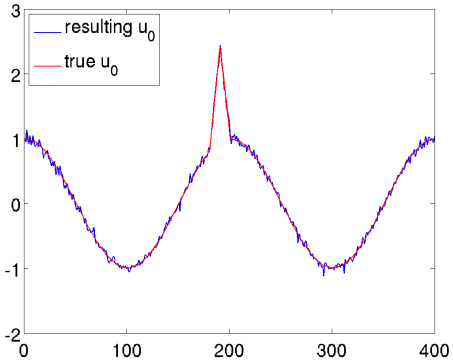


(b) True $u(0)$
and algorithm solution

Figure 6.4: Results: Nonlinear 1D Wave equation

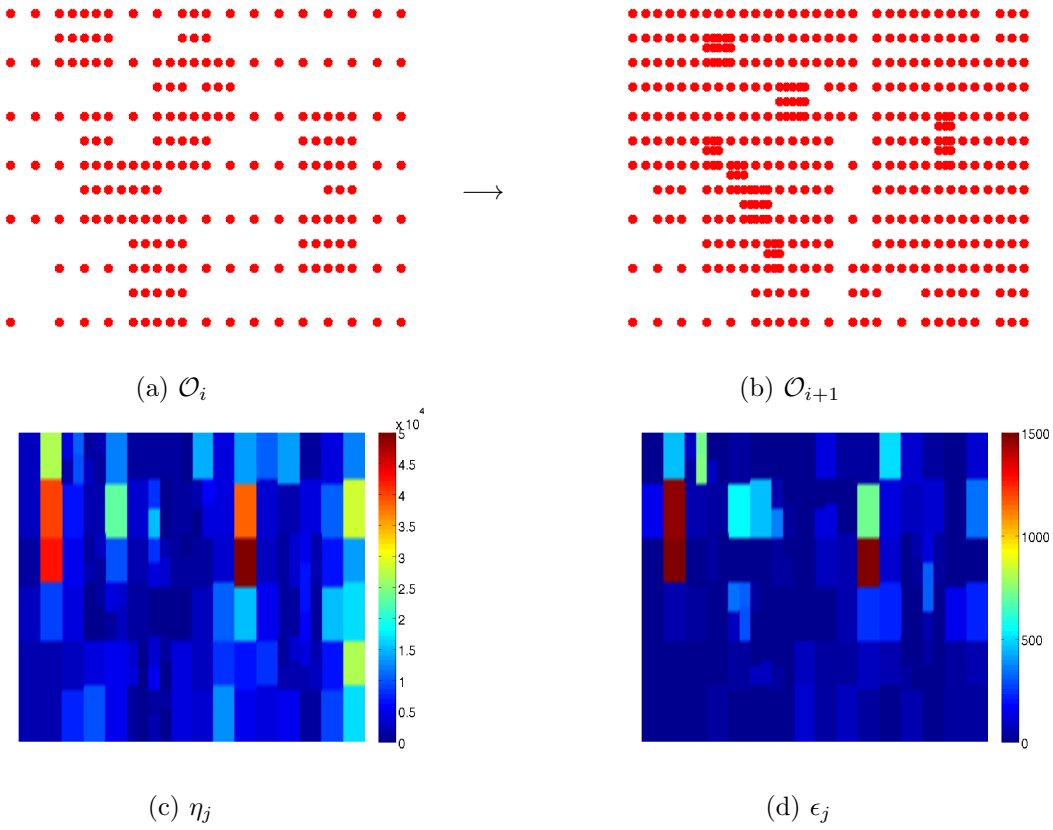


(a) Background vector
and true initial $u(0)$



(b) True $u(0)$
and algorithm solution

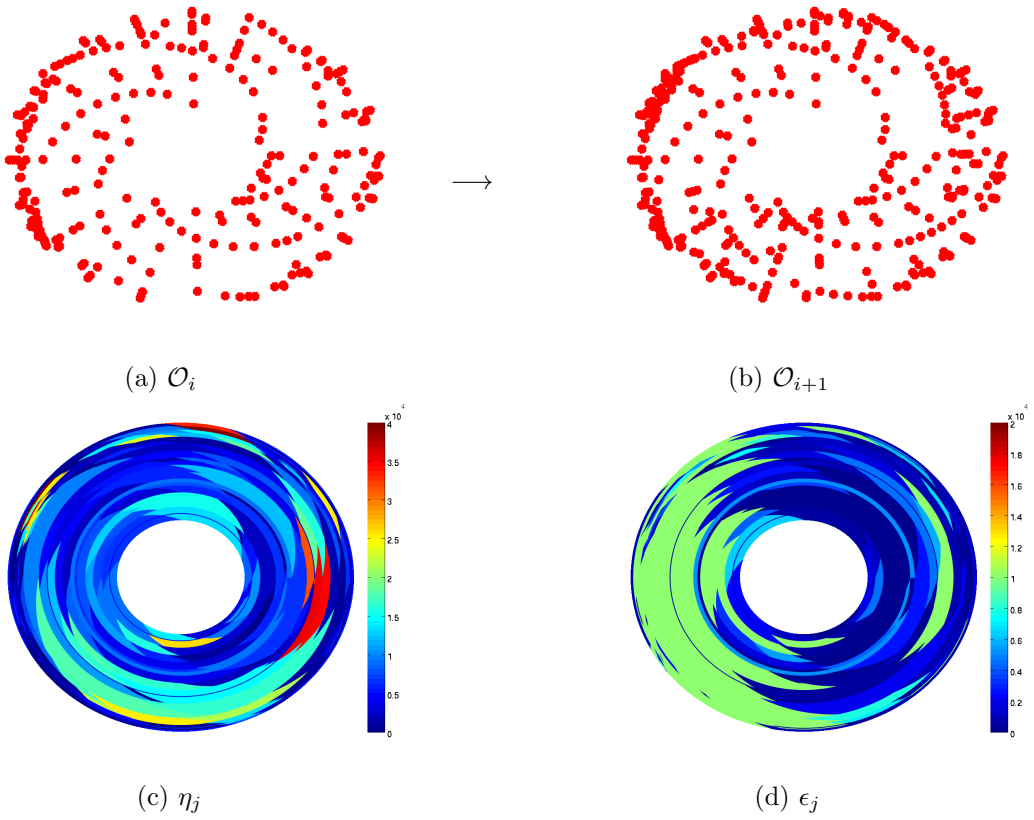
Figure 6.5: Results: Lorenz96



(a) $\mathcal{O}_i$ (b) $\mathcal{O}_{i+1}$

(c) η_j (d) ϵ_j

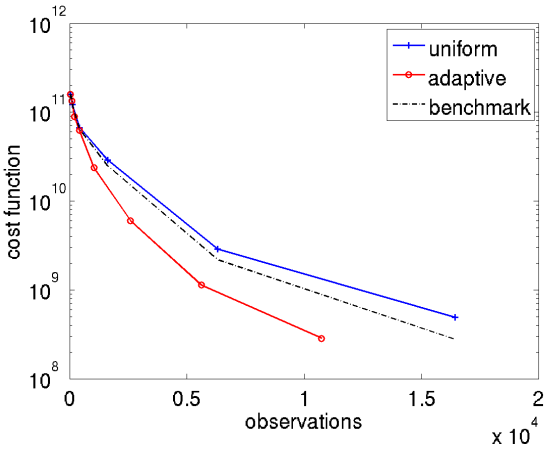
Figure 6.6: Observations set and adaptive errors



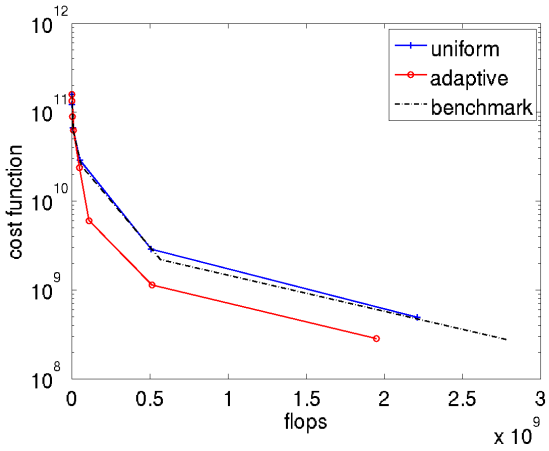
(a) $\mathcal{O}_i$ (b) $\mathcal{O}_{i+1}$

(c) η_j (d) ϵ_j

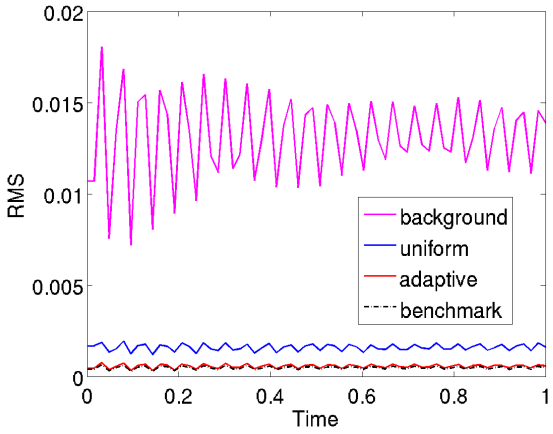
Figure 6.7: Observations set and adaptive errors



(a) Cost function versus number of observations

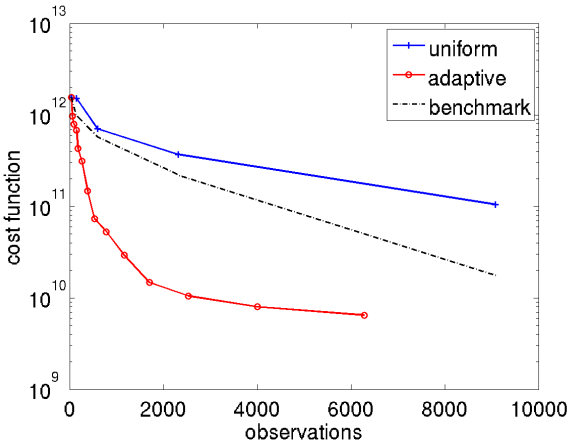


(b) Cost function versus flops

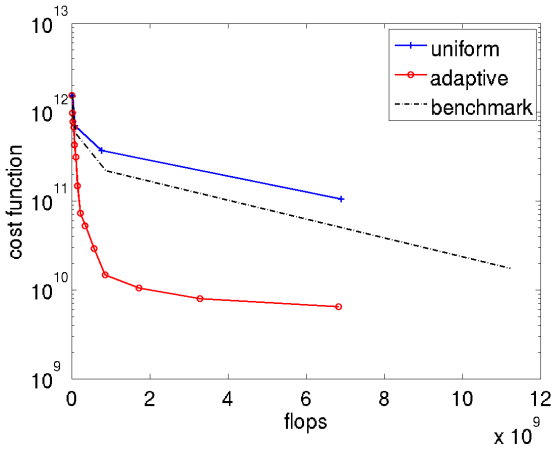


(c) Optimized solution : RMS error versus time

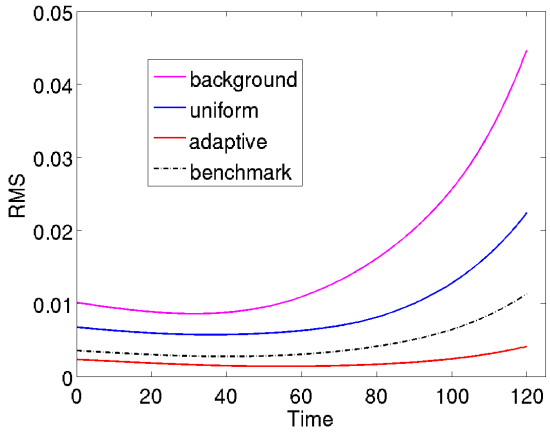
Figure 6.8: Performance of the algorithm on the nonlinear wave equation



(a) Cost function versus number of observations



(b) Cost function versus flops



(c) Optimized solution : RMS error versus time

Figure 6.9: Performance of the algorithm on the Lorenz96 problem

References

M. Arioli. A stopping criterion for the conjugate gradient algorithm in a finite element method framework. *Numerische Mathematik*, **97**(1), 1–24, 2004.

A. Brandt. Multi-level adaptive technique (MLAT) for fast numerical solution to boundary value problems. *in* 'Proceedings of the Third International Conference on Numerical Methods in Fluid Mechanics', pp. 82–89. Springer, 1973.

C. Cardinali, S. Pezzulli, and E. Andersson. Influence-matrix diagnostic of a data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **130**(603), 2767–2786, 2004.

P. Courtier. Dual formulation of four-dimensional variational assimilation. *Quarterly Journal of the Royal Meteorological Society*, **123**(544), 2449–2461, 1997.

P. Courtier, J.N. Thépaut, and A. Hollingsworth. A strategy for operational implementation of 4D-Var, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, **120**(519), 1367–1387, 1994.

D.N. Daescu and I.M. Navon. Adaptive observations in the context of 4D-Var data assimilation. *Meteorology and Atmospheric Physics*, **85**(4), 205–226, 2004.

W. Dörfler. A convergent adaptive algorithm for Poisson's equation. *SIAM Journal on Numerical Analysis*, **33**(3), 1106–1124, 1996.

A. El Akkroui, P. Gauthier, S. Pellerin, and S. Buis. Intercomparison of the primal and dual formulations of variational data assimilation. *Quarterly Journal of the Royal Meteoralogical Society*, **134**(633), 1015–1025, 2008.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, third edn, 1996.

S. Gratton and J. Tshimanga. An observation-space formulation of variational assimilation using a restricted preconditioned conjugates gradient algorithm. *Quarterly Journal of the Royal Meteorological Society*, **135**(643), 1573–1585, 2009.

S. Gratton, S. Gürol, and Ph.L. Toint. Preconditioning and globalizing conjugate gradients in dual space for quadratically penalized nonlinear-least squares problems. *Computational Optimization and Applications*, pp. 1–25, 2010.

S. Gratton, A.S. Lawless, and N.K. Nichols. Approximate Gauss-Newton methods for nonlinear least squares problems. *SIAM Journal on Optimization*, **18**(1), 106–132, 2007.

M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of the National Bureau of Standards*, **49**, 409–436, 1952.

A. Karimi and M.R. Paul. Extensive chaos in the Lorenz-96 model. *Arxiv preprint arXiv:0906.3496*, 2009.

A. Logg, K.A. Mardal, and G.N. Wells. Automated solution of differential equations by the finite element method. *Lecture Notes in Computational Science and Engineering*, **84**, 1–736, 2012.

E.N. Lorenz and K.A. Emanuel. Optimal sites for supplementary weather observations: Simulation with a small model. *Journal of the Atmospheric Sciences*, **55**(3), 399–414, 1998.

S. McCormick. Fast adaptive composite grid (FAC) methods: Theory for the variational case. *in* 'Defect correction methods', pp. 115–121. Springer, 1984.

P. Morin, R.H. Nochetto, and K.G. Siebert. Data oscillation and convergence of adaptive fem. *SIAM Journal on Numerical Analysis*, **38**(2), 466–488, 2000.

Nocedal and Wright. Theory of constrained optimization. *in* 'Numerical Optimization', pp. 314–357. Springer, 1999.

M. M. Rincon-Camacho. *Adaptive Methods for Total Variation Based Image Restoration.* Ph.D. Thesis, University of Graz, 2011.